PUAF 610
DISCUSSION SECTION 11

Wednesday, November 14, 2012
Mariel Borowitz
mariel@umd.edu
Plan for Today

- Issues from Problem Set 8
- Lecture 11 Review
- Discuss Readings
- Problem Set 9
Issues from Problem Set 8

• Significant Digits!
• Interpreting Chi-square Test Results
  • Meaning of the p-value
  • Whether to reject null hypothesis or not
  • What the null hypothesis says (and what it means if you reject it)
• Assumptions for Chi-square Testing
  • If more than 20% of your values are under 5 (or 10% are under 1),
    you need to combine categories and/or use the Fisher Exact Test
Lecture 11 Review: Regression Analysis

- Regression Analysis vs. Chi-square and Hypothesis testing
- Least Squares Estimation
- Linear Model
- Error and Homoscedasticity
- Residuals and Standard Error
- Goodness of Fit
- R-squared
- Inferences about Slope
- Inferences about Correlation
- Excel Functions
Regression Analysis vs. Chi-square and Hypothesis Testing

• What do we compare when doing a hypothesis test?
  • One mean (or proportion) to a population mean (or proportion), OR
  • Two means (or proportions) to each other

• What is an example of a null hypothesis and alternative hypothesis when carrying out a hypothesis test?
  • \( H_0: \mu_1 = \mu_2 \) (i.e., the mean of the two samples are equal, i.e. they come from the same population)
  • \( H_A: \mu_1 \neq \mu_2 \) (i.e., the mean of the two samples are not equal, i.e. they do not come from the same population)
  • This example was two-sample (but not matched pairs) & two-tailed
Regression Analysis vs. Chi-square and Hypothesis Testing

- What do we compare when doing a chi-square test?
  - One variable (with multiple values) to the expected distribution of that variable, OR
  - Two variables (with multiple values) to each other

- What is an example of a null hypothesis and alternative hypothesis when carrying out a chi-square test?
  - $H_0$: The two categorical variables are independent
  - $H_A$: The two categorical variables are not independent; i.e. an association or correlation exists between the two variables
  - This example was for a chi-square test for independence

- How can we find out the strength of the relationship between two variables?
  - Use regression analysis

- How can we predict one variable based on a known value of the other?
  - Use regression analysis
Regression Analysis vs. Chi-square and Hypothesis Testing

- What do we compare when doing a simple (bivariate) regression analysis?
  - Two variables (with multiple values) to each other

- What output of the regression can be used to predict values of one variable based on the other?
  - The regression equation \(y=bx+a\)

- What output of the regression can be used to determine the strength and direction of the relationship between two variables?
  - The correlation coefficient

- What output of the regression can be used to determine the proportion of the variability that is explained by the regression?
  - The \(R^2\) value
Least Squares Estimation

• Start with a collection of data points – two data points for each observation \((x, y)\)
• Try to come up with a way to describe the relationship between the \(x\) and \(y\) variables. (Can also think of this as a way to predict \(y\), given \(x\).)
Least Squares Estimation

- The mean value of y tells us something about the y-value we expect given an x-value
- First way to describe y: the average value of y for every x

Average y value is 30.2
Can also think of this as a line with slope of zero, intercept of 30.2
Least Squares Estimation

- Still a lot of unexplained variation using this estimator
- Thin, gray lines show variation of observed y-values from the mean value of y (y-bar)

Average y value (y-bar) is 30.2
Least Squares Estimation

- We would like to have as little gray on this chart as possible
- i.e. we would like to minimize the sum of the residuals
Least Squares Estimation

• Draw a new line that makes the gray lines (the residual error) as small as possible
• But we can’t just do this by hand, we need math
Least Squares Estimation

- The equation of a line is: \( y = mx + b \), where \( m \) is the slope and \( b \) is the y-intercept (where the line crosses the y-axis)
  - In regression analysis, we use \( \beta \) (or \( b \)) instead of \( m \) and \( \alpha \) (or \( a \)) instead of \( b \)
  - So, we want to find a line (define \( a \) and \( b \)) that will minimize the total length of all the gray lines (minimize the amount of error overall)

Want to find a line:

\[ y = a + bx \]

Where the total amount of gray is minimized

So we need to find out the correct values of \( a \) and \( b \)
Least Squares Estimation

- However, since some are of the errors are positive and some are negative, if we just add them all together, some will cancel out.
- Solution: Square those values before we add them together.
- Minimizing the square of the errors (variation from the predicted line) is least squares estimation!

\[
y = a + bx
\]

Calculate \(a\) and \(b\) to find the line that minimizes the square of the residuals.

In this case:
\[
y = 1.6x + 12
\]
Least Squares Estimation: The Math

- What does it mean to minimize the sum of the residuals squared?
  - Minimize the sum of the squared residuals \( \sum_{i=1}^{n} e_i^2 \)
  - Same as minimizing the sum of the square of the differences between each observed \( y \) and predicted \( y \)
  - Same as minimizing the sum of the \((y_i - a - bx_i)^2\)
    - Get this by rearranging the equation on the right and squaring

- Minimizing requires taking the partial derivatives of the last equation \((y_i - a - bx_i)^2\) with respect to \( a \) and with respect to \( b \) and setting them each to zero (since a derivative gives you the slope, and a slope of zero is a minimum or maximum)

- For more on this…
  - To read through a good step-by-step description of the derivation:
    - http://isites.harvard.edu/fs/docs/icb.topic515975.files/OLSDerivation.pdf
  - Or to watch some multivariable calculus being done (27:17 for OLS):
Least Squares Estimation: The Math

• After you do the calculations, you get this:

\[ \beta \equiv b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{s_x^2} = r \frac{s_y}{s_x} \]

\[ \alpha \equiv a = \bar{y} - b\bar{x} \]
The Linear Model

• What do \(x\), \(y\), \(\alpha\), and \(\beta\) represent in the following equation?

\[
y = \alpha + \beta x + \epsilon
\]

• \(x\) = independent or explanatory variable (“cause”)
• \(y\) = dependent or response variable (“effect”)
• \(\alpha\) = intercept (value of \(y\) when \(x = 0\))
• \(\beta\) = slope (change in \(y\) when \(x\) increases one unit)
• \(\epsilon\) is normally distributed random variable, with a mean of 0 and a standard deviation of \(\sigma\)

• What are \(y\)-bar, \(y_i\), and \(\hat{y}_i\)?
  • \(y\)-bar is the mean value of \(y\)
  • \(y_i\) is a particular value of \(y\) (from the i-th observation)
  • \(\hat{y}_i\) is the expected value for a particular \(y\) (from the i-th observation)
Error and Homoscedasticity

- Why do we include an error term in the linear model?
  - Linear relationships are rarely precise because of measurement error and inherent variability
  - Relationships with zero error are usually trivial (i.e. relationship between degrees Fahrenheit and degrees Celsius)

- What is the definition of homoscedasticity?
  - The standard deviation of the error is constant and independent of x (we’ll talk more about this – and the alternative – next week)
Residuals and Standard Error

• What key assumptions do we make about the residuals?
  • Residuals are independent
  • Residuals are normally distributed
  • Residuals have a constant standard deviation

\[ \sigma \approx s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n - 2}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n - 2}} = \sqrt{\frac{\text{SSE}}{n - 2}} \]
Goodness of Fit

- What does SST represent?
  - The total variability in y (sum of squares, total)
- What are the two components of SST?
  - SST = SSR + SSE
  - SSR: Variability explained by the regression (sum of squares, regression)
  - SSE: Remaining or unexplained variability (sum of squares, error)
- Explain the following equation in words (hint – use the equivalent equation on the bottom)

\[
\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2
\]

- The sum of the distance of each observed y from the mean y is equal to the sum of the square of the distance from each expected y to the mean y (the variability explained by the regression) added to the sum of the square of the distance from the observed y to the expected y (the remaining unexplained variability)
Goodness of Fit

- Which of these diagrams shows total error? Explained error (due to regression)? Unexplained or residual error?
R-squared

• What does R-squared tell us? What is another name for this term?
  • R-squared tells us the proportion of the variability that is explained by the regression
  • It is also known as the “coefficient of determination”

• What does an R-squared of one imply? When might you see this?
  • R-squared of one implies perfect linear correlation
    • Example: Degrees Celsius and Degrees Fahrenheit

• What does an R-squared of zero imply?
  • There is no correlation between the variables; they are independent
R-squared

- How can you find R-squared in terms of SSR, SST, and/or SSE? Give the answer as an equation and in words.

\[
R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \left[ \frac{\text{cov}(x, y)}{s_x s_y} \right]^2 = 1 - \left( \frac{s_e}{s_y} \right)^2
\]

- R-squared is the variability explained by the regression (sum of squares, regression) divided by the total variability (sum of the square, total)
- i.e., R-squared is the proportion of the the variability that is explained by the regression
Correlation and Causation

- Are each of these pairs of variables likely to be correlated? Would it be reasonable to believe there is causation?
  - Number of people per day wearing sunglasses and number of people per day getting sunburned
  - Smoking cigarettes and rates of lung cancer
  - Cell phone use and brain cancer
  - Eating breakfast and doing well in school
  - Gun ownership and crime

- What could you want to do to be more sure about causation?
  - Run an experiment with a control group
Correlation and Causation

• Does correlation prove correlation?
  • No!

• To quote xkcd: Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.
Inferences about Slope

- It is possible to carry out a hypothesis test to determine the likelihood that the variables are independent.
- What is the null hypothesis and alternative hypothesis for this test?
  - \( H_0 \): Slope = 0 (i.e. no association between x and y)
  - \( H_A \): Slope \( \neq \) 0 (i.e. correlation between x and y)

- How do you find the t-statistic for this test? How do you find the degrees of freedom? How do you find the p value?
  
  \[
  s_b = \frac{1}{\sqrt{n-1}} \frac{s_e}{s_x} \\
  \text{df} = n - 2 \text{ for one independent variable} \\
  t = \frac{b}{s_b} \\
  \text{T.Dist}(t\text{-value}, \text{df})
  \]
Inferences about Correlation

- There is another way to carry out a hypothesis test to determine the likelihood that the variables are independent, using just R-squared and n.
- What is the null hypothesis and alternative hypothesis for this test?
  - \( H_0: \text{R-squared} = 0 \) (i.e. no association between year and temperature)
  - \( H_A: \text{R-squared} \neq 0 \) (i.e. correlation between x and y)

- How do you find the t-statistic for this test? How do you find the p value?
  \[
  s_r = \sqrt{\frac{1 - r^2}{n - 2}} \\
  t = \frac{b}{s_b} = \frac{r}{s_r} = \sqrt{\frac{(n - 2)R^2}{(n - 2) - 1 - R^2}}
  \]
  \[T.DIST(t\text{-value, df})\]
Excel Functions

- Scatterplot:
  - Chart/Add Trendline/Linear/Display equation, R2

- Menu-driven tools:
  - Tools/Data Analysis/Regression

- Excel functions:
  - $a = \text{INTERCEPT}([y\text{-range}],[x\text{-range}])$
  - $b = \text{SLOPE}([y\text{-range}],[x\text{-range}])$
  - $se = \text{STETYX}()$
  - $\text{LINEST}([y\text{-range}],[x\text{-range}],1,1)$ gives all of above

- Y-hat (expected y) = $\text{FORECAST} (x,[y\text{-range}],[x\text{-range}])$
Readings

• Carlberg – Chapter 4: How Variables Move Jointly: Correlation
• Kahneman – Chapter 17-18
Understanding Correlation

The correlation coefficient, $r$, expresses the strength and direction of the relationship between two variables.

- $r$ always varies from -1 to 1.
- Negative values imply a negative correlation.
- Positive values imply a positive correlation.
- The closer $r$ is to plus or minus 1, the stronger the relationship.
- When two variables are unrelated, the correlation between the two of them is zero (or close to zero).

CORREL([x-values],[y-values])
Understanding Correlation

- Equation for variance was covered in lecture 2
  - Subtract the mean from each value and square the deviation
  - Sum of \((X_i - \bar{X})(X_i - \bar{X})\) for all values of \(x\)
- Correlation is based on covariance \((s_{xy})\)
  - Same idea as variance for one variable discussed in early lectures, but now you take the deviation score from one variable and multiply it by the deviation score for the other variable

\[
s_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})
\]
Understanding Correlation

- Imagine you have a dataset showing height (x) and weight (y)
  - The mean height is 67 inches
  - The mean weight is 175 pounds

- What happens to the numerator of the covariance equation if you have a person is 72 inches tall and that weighs 150 pounds (both values above average)?
  - \((X_i - X_{\text{bar}})\) is positive and \((Y_i - Y_{\text{bar}})\) is negative
  - \((X_i - X_{\text{bar}})(Y_i - Y_{\text{bar}})\) is negative

- What happens if you have a person that is 62 inches tall and weights 200 pounds (both values below average)?
  - \((X_i - X_{\text{bar}})\) is negative and \((Y_i - Y_{\text{bar}})\) is positive
  - \((X_i - X_{\text{bar}})(Y_i - Y_{\text{bar}})\) is negative

- These two negative numbers add together in the covariance equation to make a bigger negative number (farther from zero)
**Understanding Correlation**

- Imagine you have a dataset showing height (x) and weight (y)
  - The mean height is 67 inches
  - The mean weight is 175 pounds

- What happens to the numerator of the covariance equation if you have a person that is 72 inches tall and weighs 200 pounds (both values above average)?
  - \((X_i - \bar{X})\) is positive and \((Y_i - \bar{Y})\) is positive
  - \((X_i - X-bar)*(Y_i - Y-bar)\) is positive

- What happens if you have a person that is 62 inches tall and weights 150 pounds (both values below average)?
  - \((X_i - X-bar)\) is negative and \((Y_i - Y-bar)\) is negative
  - \((X_i - X-bar)*(Y_i - Y-bar)\) is positive

- These add together in the covariance equation to make a larger positive number (farther from zero in the positive direction)

\[
s_{xy} = \sum_{i=1}^{N} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(N - 1)}
\]
Understanding Correlation

• What happens to the previous example if you add someone that is 75 inches tall and weighs 170 pounds?
  • (Xi-X-bar) is positive and (Yi-Y-bar) is negative
  • (Xi – X-bar)*(Yi – Y-bar) is negative
  • Add this negative number to the other two positive numbers, and you bring the covariance back towards zero (a smaller co-variance).

• If the trend for the points is in the same direction (all positive correlation or all negative), the relationship between the two variables will be stronger
• When the trend is different for different points, the relationship will be weaker
• Remember: this is not the only determinant of strength!
Understanding Correlation

• Moving from Covariance to the Correlation
• The correlation is equal to the covariance divided by the product of the standard deviation of $x$ and the standard deviation of $y$

$$r = \frac{s_{xy}}{s_x s_y}$$

• This normalizes the correlation – it ensures that it ranges only between -1 and 1
• Bust since the covariance is in the numerator, things that make the covariance larger (as shown in the previous slides), will make the correlation larger, and vice versa
Correlation and R-squared

• From the book:

\[ s_{xy} = \sum_{i=1}^{N} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{(N - 1)} \]

\[ r = \frac{s_{xy}}{s_x s_y} \]

\[ r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} \]

• From lecture:

\[ r = \frac{\text{cov}(x, y)}{s_x s_y} = b \frac{s_x}{s_y} \]

\[ R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = \left[ \frac{\text{cov}(x, y)}{s_x s_y} \right]^2 = 1 - \left( \frac{s_e}{s_y} \right)^2 \]

• What output of the regression can be used to determine the strength and direction of the relationship between two variables?
  • The correlation coefficient

• What output of the regression can be used to determine the proportion of the variability that is explained by the regression?
  • The \( R^2 \) value
Chapter 17

• Israeli Air Force Instructor:
• “On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver. The next time they try the same maneuver they usually do worse. On the other hand, I have often screamed into a cadet’s earphone for bad execution, and in general he does better on his next try. So please don’t tell us that reward works and punishment does not, because the opposite is the case.”
Chapter 17: Example One

• Assuming that the observations made by the instructor are correct, why might his conclusion that negative reinforcement works better than positive reinforcement still be wrong?
  • Regression to the Mean!
  • Results are due partly to skill and partly to luck (variance)
  • “Naturally, he praised only a cadet whose performance was far better than average. But the cadet was probably just lucky on that particular attempt and therefore likely to deteriorate regardless of whether or not he was praised. Similarly, the instructor would shout into a cadet’s earphones only when the cadet’s performance was unusually bad and therefore likely to improve regardless of what the instructor did.”
Chapter 17: Example Two

• Think of a two-day golf tournament where the average score of the golfers was a par 72. Assume Golfer1 scored 66 and Golfer2 scored 77. How would you expect each of them to do in the second day?
  • Their scores are due to talent and/or luck.
  • Above average on day 1 = above-average talent + lucky on day 1
  • Below average on day 1 = below-average talent + unlucky on day 1
  • We expect their talent to stay the same on day two, but we have no way to predict luck – so we have to assume they’ll have an average day in terms of luckiness
  • “The golfer who did well on day 1 is likely to be successful on day 2 as well, but less than on the first, because the unusual luck he probably enjoyed on day 1 is unlikely to hold.”
  • “The golfer who did poorly on day 1 will probably be below average on day 2, but will improve, because his probable streak of bad luck is not likely to continue.”
Chapter 17: Example Three

- You may have heard of the “Sports Illustrated jinx”: “the claim that an athlete whose picture appears on the cover of the magazine is doomed to perform poorly the following season.” What reasons are typically given to explain this? What reason might you give as a statistician?
  - Typical Explanations:
    - Overconfidence
    - Pressure of meeting high expectations
  - Statistician explanation (simpler and more reasonable):
    - “An athlete who gets to be on the cover of Sports Illustrated must have performed exceptionally well in the preceding season, probably with the assistance of a nudge from luck— and luck is fickle.”
Chapter 17: Regression to the Mean

- What does regression to the mean have to do with regression analysis?
  - The connection has to do with the idea of correlation
  - Regression to the mean occurs when the correlation between variables is not perfect.
    - If your golf score today was a perfect predictor of your golf score tomorrow (they were perfectly correlated), we wouldn’t see the regression to the mean.
    - But though the two scores are related (correlated), because they are both affected by your talent level, the correlation is not perfect (in the example, this is because luck plays a role)
    - Therefore, the best prediction of your score on the second day is a score that is closer to the mean than your score the first day.
Chapter 17: Example Four

• Consider the following proposition:
• Highly intelligent women tend to marry men who are less intelligent than they are.
• How can we explain this?
  • Common answers refer to avoiding competition between spouses, or preferences of intelligent men for less intelligent women.
• What if I tell you: The correlation between the intelligence scores of spouses is less than perfect.
  • “If the correlation between the intelligence of spouses is less than perfect (and if men and women on average do not differ in intelligence), then it is a mathematical inevitability that highly intelligent women will be married to husbands who are on average less intelligent than they are (and vice versa, of course).”
Chapter 17: Example Five

• You are the sales forecaster for a department store chain. All stores are similar in size and merchandise selection, but their sales differ because of location, competition, and random factors. You are given the results for 2011 and asked to forecast sales for 2012. You have been instructed to accept the overall forecast of economists that sales will increase overall by 10%. How would you complete the following table?

<table>
<thead>
<tr>
<th>Store</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$11,000,000</td>
<td>__________</td>
</tr>
<tr>
<td>2</td>
<td>$23,000,000</td>
<td>__________</td>
</tr>
<tr>
<td>3</td>
<td>$18,000,000</td>
<td>__________</td>
</tr>
<tr>
<td>4</td>
<td>$29,000,000</td>
<td>__________</td>
</tr>
<tr>
<td>Total</td>
<td>$61,000,000</td>
<td>$67,100,000</td>
</tr>
</tbody>
</table>
Chapter 17: Example Five

• “Having read this chapter, you know that the obvious solution of adding 10% to the sales of each store is wrong.

• You want your forecasts to be regressive, which requires adding more than 10% to the low-performing branches and adding less (or even subtracting) to others.

• But if you ask other people, you are likely to encounter puzzlement: Why do you bother them with an obvious question? As Galton (who discovered/formalized the concept of regression to the mean) painfully discovered, the concept of regression is far from obvious.”
Chapter 18: Example One

• Julie is currently a senior in a state university. She read fluently when she was four years old. What is her grade point average (GPA)? What is her percentile score on GPA?

• What is Julie’s percentile score on reading precocity?
Chapter 18: Example One

• A high school counselor describes a student with the following description:
  • intelligent, self-confident, well-read, hardworking
• What percentage of descriptions of college freshman do you believe would impress you more?
• What is the percentage of freshmen who obtain higher GPAs?
Chapter 18: Example One

- If you’re like most people, you gave the same percentile score for both answers in Example One and Example Two. Why might this be a problem?

- The question about reading precocity and about the number of descriptions that would be more impressive only require you to evaluate the evidence.

- The questions about GPA require you to make a prediction, which involves a great deal of uncertainty.

- So how could we do a better job?
Chapter 18: Taming Intuitive Predictions

• Schematic formula for the factors that determine reading age and college grades:
  • reading age = shared factors + factors specific to reading age = 100%
  • GPA = shared factors + factors specific to GPA = 100%

• So there is some correlation between the two (some shared factors influencing both), but how high do you think this correlation is? Certainly not 100%, so…
  1. Start with an estimate of average GPA.
  2. Determine the GPA that matches your impression of the evidence.
  3. Estimate the correlation between your evidence and GPA.
  4. If the correlation is .30, move 30% of the distance from the average to the matching GPA.
Chapter 18: Taming Intuitive Predictions

• This is similar to the issue we looked at earlier with conditional probability
  • In those examples, we were tempted to give a probability as our answer that used the information at hand (Tom is shy and organized) and ignore the base rate probability (size of various university departments)
  • In these examples, we were tempted to give a predicted value based only on the information at hand (reading ability at age four) and ignore the other important variables needed to predict the answer (variables that affect GPA, but aren’t related to reading ability at age four)
Chapter 18: Taming Intuitive Predictions

• Furthermore, you should know that correcting your intuitions may complicate your life.

• A characteristic of unbiased predictions is that they permit the prediction of rare or extreme events only when the information is very good.

• If you expect your predictions to be of modest validity, you will never guess an outcome that is either rare or far from the mean.

• If your predictions are unbiased, you will never have the satisfying experience of correctly calling an extreme case.

• You will never be able to say, “I thought so!” when your best student in law school becomes a Supreme Court justice, or when a start-up that you thought very promising eventually becomes a major commercial success.
1. Use the data in UMCPsalaries.xlsx to do a simple linear regression analysis using two variables: salary and years since highest degree (yrdeg).

1. What relationship do you expect between these variables? Which should be considered the independent variable? Which should be plotted on the x-axis?

2. Create a scatterplot. Using the "trend line" option, display the regression line, the regression equation, and the value of R-squared on the scatterplot.

3. Explain the meaning of the slope and the intercept in plain English. Do the values make sense?

4. Explain the meaning of R-squared in plain English.

5. Now use the data analysis toolpack to perform the regression analysis and verify that your results for the slope, intercept, and R-squared are the same as given above. (If you are a Mac user, copy the data into reg.xlsx and enter the name of the data tab and the appropriate column and row references.)

6. Formulate and test an appropriate null hypothesis for the slope. What do you conclude?

7. What is the standard error of the residuals? Explain the meaning of this number in plain English.

8. Plot the residuals. What do you notice? What is this pattern called? Why does it arise?
Problem Set 9

2. Use the file SmokingCancer.xlsx contains three datasets on smoking and cancer. Use one of the first two datasets to do a simple linear regression analysis of the relationship between smoking behavior and the incidence of lung cancer.
   1. Create a scatterplot and display the regression line, the regression equation, and the value of R-squared on the scatterplot.
   2. Explain the meaning of the slope and the intercept in plain English.
   3. Explain the meaning of R-squared in plain English.
   4. Formulate and test an appropriate null hypothesis for the slope. What do you conclude?

3. The file RMBL.xlsx contains data collected by UMCP Prof. David Inouye (and his graduate students) at the Rocky Mountain Biological Laboratory over the past 38 years.
   1. Is the mean minimum temperature in April increasing? Is this increase statistically significant? Test the appropriate null hypothesis.
   2. Do the same for the date of first bare ground. What do you conclude?
   3. Explain the meaning of the slopes from parts (2) and (3) in plain English.
   4. One might hypothesize that the date marmots come out of hibernation (as measured by the date of first sighting) is related to climate. Is the date of first sighting better explained by the mean minimum April temperature, or by the date of first bare ground?
Questions?